Heuristic best-first search in separation of interleaved Web sessions

Matjaž Kukar
matjaz.kukar@fri.uni-lj.si

University *of Ljubljana*
Faculty of *Computer and Information Science*

Overview

- Introduction
- Clickstream data
- Motivation: interleaved sessions
- Separation process with best-first search
- Evaluation methods
- Data: university student records IS, large web shop
- Results
- Conclusion

Introduction

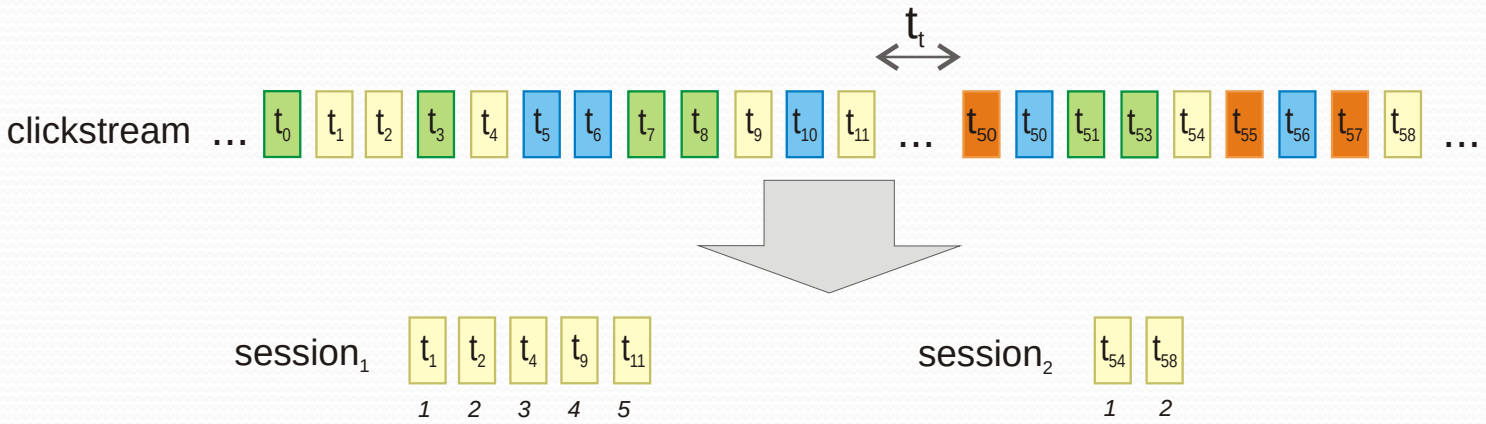
- Web sites important for companies
  - Sell products, services, data access
- Strong competition
- Web pages
  - Complexity of sites rising
  - Increased number of users
- Importance of web site visitor's behavior
  - Customizing pages → better user experience

Clickstream

- Main source of data for user behavior analysis
- Clickstream
  - A sequence of clicks user makes
  - Detailed view on user transitions between pages
  - Source: HTTP server log file (CLF, ECLF)
- Incomplete picture of user's activity
  - Noisy, large, duplicated data
  - Inadequately structured
  - No user session is logged
- Needs to be preprocessed and cleaned
  - Sessionization – gather all individual events
  - Hard to reliably identify user sessions

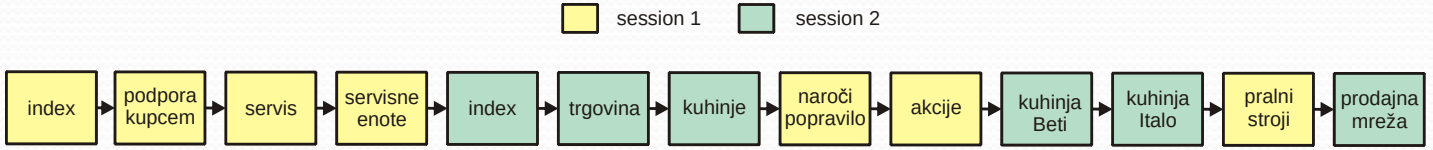
User session

- User session
  - One visit of a user to a web site
  - In order to do one or more tasks
- Sessionization prone to errors
- The problem of interleaved sessions



Interleaved sessions

- User session with interleaved actions from several browser windows/tabs
  - A single long user session
    - Consists of two or more sesions
  - Conceals actual user intentions
- Reasons
  - Parallel user behavior
  - Users often browse the same site:
    - With multiple browsers opened, multiple tabs
    - Switching between tasks
  - Advanced users



Effects of interleaved sessions

- Negative effect on data quality
  - … and user behaviour analyses
- Three choices
  1. Ignore the problem
     - Possibly adverse effect on data quality if too many
  2. Detect and ignore such sessions
     - Possibly discard data about valuable users
  3. Properly separate interleaved sessions
     - Cannot be easily separated
     - Context help needed

Some facts

- Student records IS
  - All interleaved sessions belong to either professors or administrators
- Web shop
  - There are twice as many buyers in interleaved sessions than in non-interleaved ones

Some combinatorics …

- Interleaving two sessions s_1 and s_2 of lengths n_1 and n_2:

$$\binom{n_1 + n_2}{n_1} = \binom{n_1 + n_2}{n_2}$$

- Sessions as sets? What about the order of elements?

Some combinatorics …

- Interleaved session *s* of length *n*: in how many ways can it be constructed from up to *n* non-empty sessions?
- Bell numbers: the number of ways a set of *n* elements can be partitioned into (up to *n*) nonempty subsets

$$B_n = \sum_{k=1}^{n-1} \binom{n}{k} B_{n-1} \qquad B_0 = 1$$

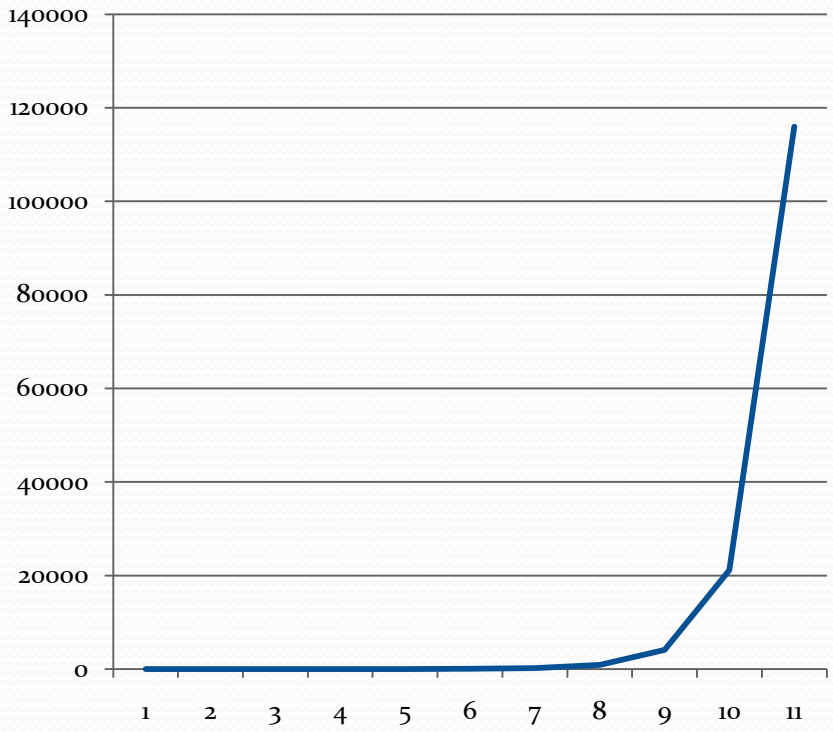
Some combinatorics …

- Interleaved session s of length n: in how many ways can it be constructed from exactly k sessions?
- Stirling numbers of the second kind: the number of ways of partitioning a set of n elements into k nonempty sets

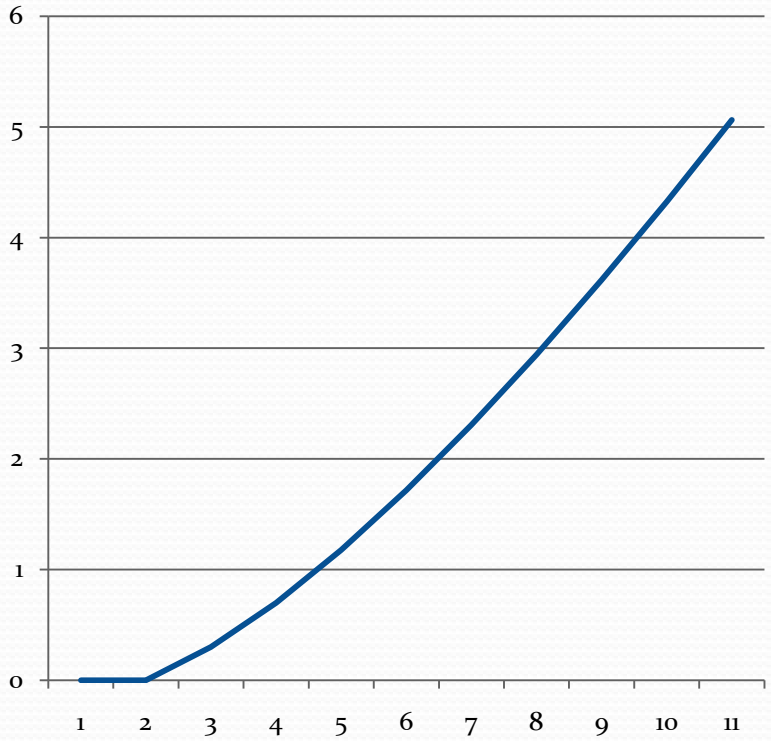
$$S_{n,k} = \left\{ {n \atop k} \right\} = \frac{1}{k!} \sum_{i=0}^{k} (-1)^i \binom{k}{i} (k-i)^n$$

- Unsurprisingly $B_n = \sum_{k=0}^{n} S_{n,k}$

Bell numbers
(number of possible separations)



Linear scale



Log scale

Separating method

- Discrete Markov model (MM) used for data representation
  - Clickstream represented with first-order MM
  - Present users' path through web site
  - Model trained (probabilities) with validated clean sessions
- Training MM
  - Background knowledge
  - Training data
- Separation based on former user behavior
  - Searching in state space
  - Uses trained Markov model

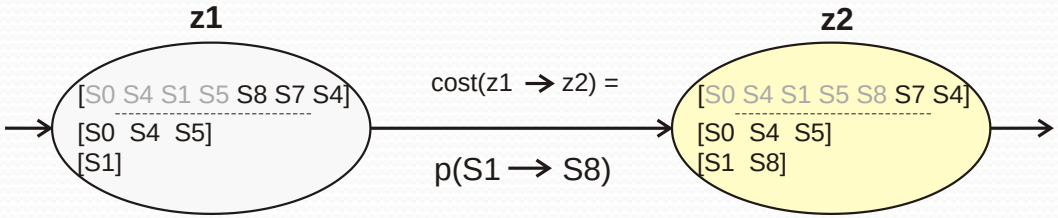
Separating with state space search

- Problem of separation transformed into the problem of searching alternatives in MM state space
- State Z:
  - Partially separated interleaved session
  - $Z = [[S_{R_1}, S_{R_2}, \ldots S_{R_3}], S_n]$
  - $Z_S = [\ [],\ (s_1, s_2, \ldots, s_n)\ ]$
  - $Z_G = [\ [(s_{r1_1}, s_{r1_2}, \ldots s_{r1_a}), \ldots (s_{rr_1}, s_{rr_2}, \ldots s_{rr_b})], ()\ ]$

Partially separated interleaved session

Partially reconstructed sessions

- Transition between states
  - Assignment of page s_i from interleaved part
  - Starting a new separated session S_{R+1}



Separating with state space search

- Transition between states $z_1 \rightarrow z_2$ at a cost $c(z1, z2)$
  - Transition probability to page s_i
  - Start a new session (probability that s_i is a starting page)
- Probability of separated session S_R

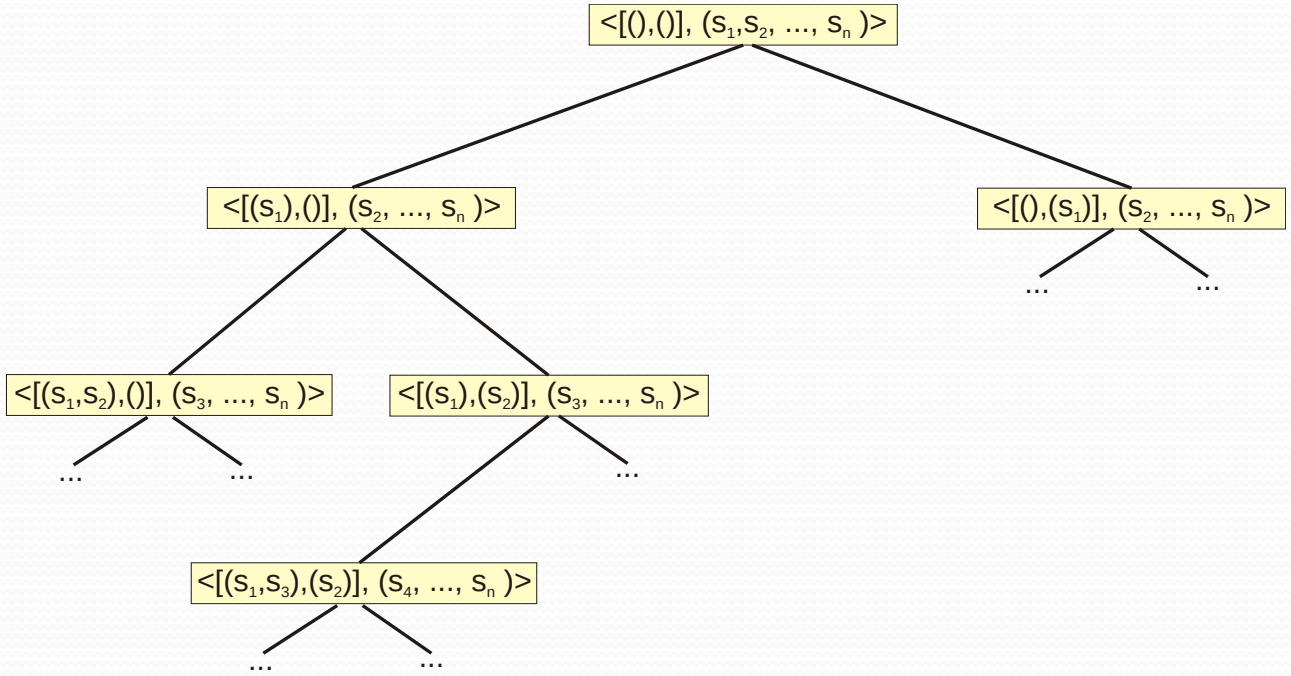
$$P(S_R) = P_{Z_S}(sr_1) \prod_{i=1}^{a-1} P(sr_i \longrightarrow sr_{i+1})$$

$$f(Z) = \prod_{i=1}^{r} P(S_{R(i)})$$

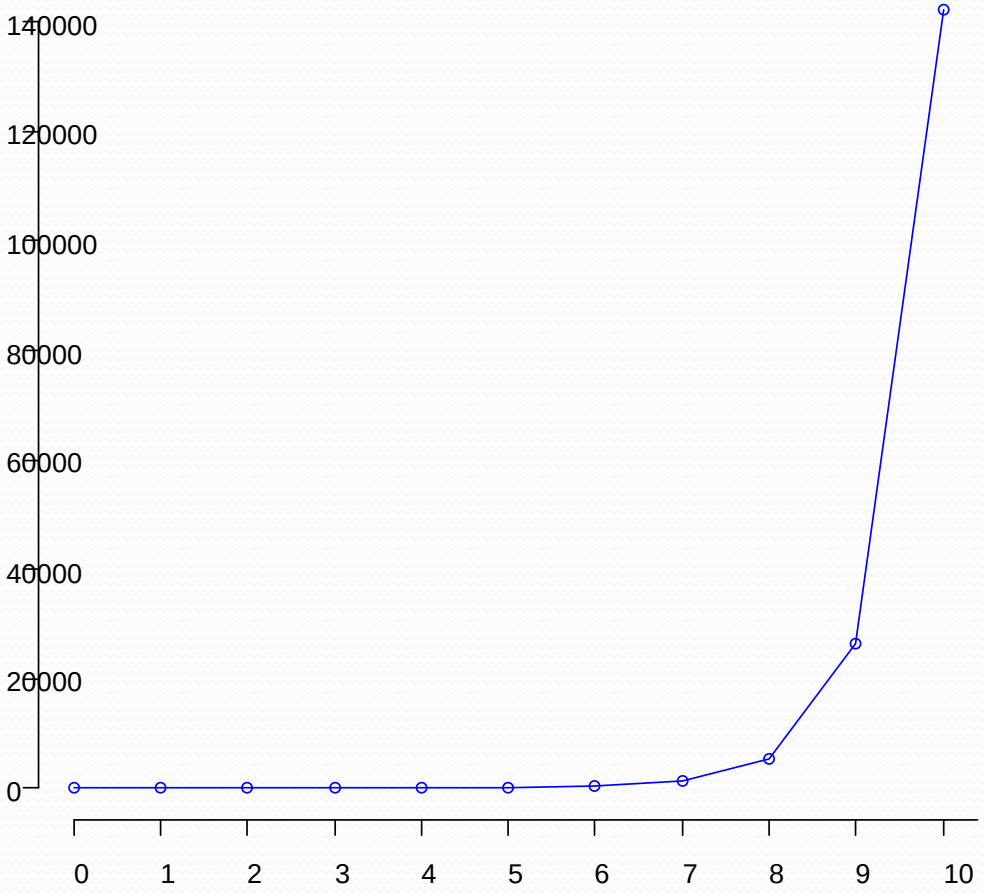
- Goal:
  - find the cheapest way between Z_S and Z_G
  - results in the most probable separation

State space

- Directed graph with actions
  - Nodes correspond to problem situations
- Number of states by level increases rapidly
  - Solution: use of heuristic search algorithm
- Sample state space limited to 2 possible separations

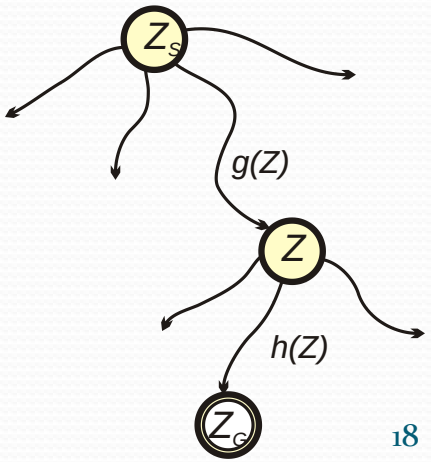


State space


140000
120000
100000
80000
60000
40000
20000
0
0
1
2
3
4
5
6
7
8
9
10

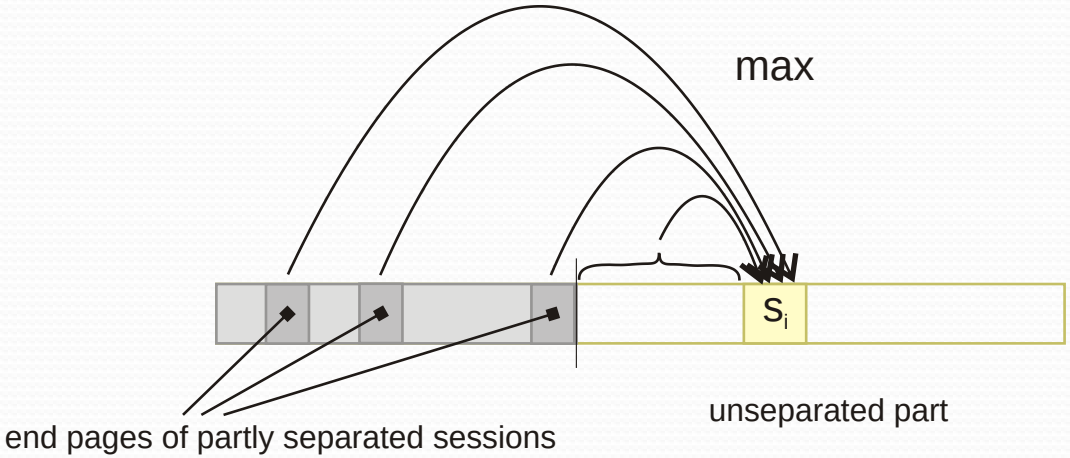
Heuristic best first search

- Potentially lower combinatorial complexity
  - Searching in direction of the most promising node
- Estimator $f(Z)$
  - $f(Z) = g(Z) + h(Z)$
    - $g(Z)$ – cost of optimal path from node Z_S to node Z
    - $h(Z)$ – estimate of the best path from node Z to goal Z_G
- Algorithm RBFS
  - Linear space complexity $O(bd)$
  - Efficient admissible heuristic function



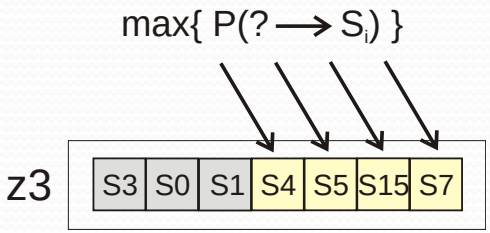
Devising a heuristic function

- Trivial heuristic function: h(Z) = 1
- Improvements, we consider
  - max probability of transition to page S_i – *max* $p(? \rightarrow S_i)$
  - Structure of session, only possible transitions
  - Transitions only from end states of partial separations



Admissibility of heuristic function

- Admissibility
  - Desired property
  - Has to optimistically estimate the nodes
  - Guarantees to find an optimal solution
- Admissible heuristics *h(z)* guarantees the most probable separation
- The most probable separation is not necessary correct solution to the problem
  - Example: interleaved sessions with low probability
- Illustration (admissibility)



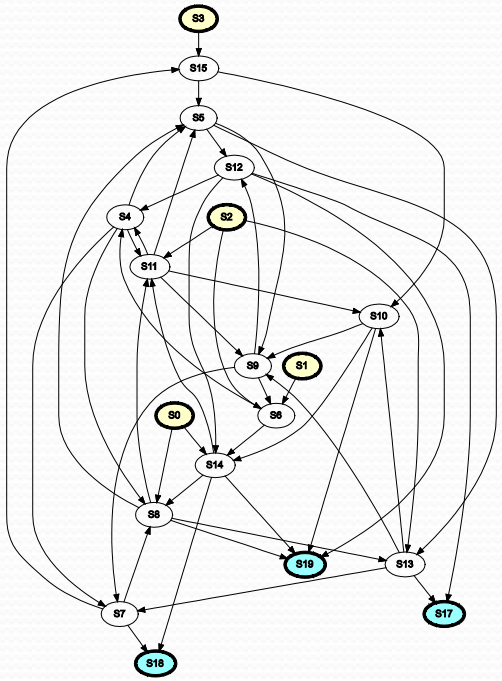
P(S0→S4) P(S1→S5) P(S5→S15) P(S4→S7) max{P(?→S4)} max{P(?→S5)} max{P(?→S15)} max{P(?→S7)}

Evaluation of separating process

- Quality of separated sessions - their similarity to original ones
- Measuring similarity between sequences – many methods.
- Methods used:
  - Perfect match
  - Similarity based on edit distance
  - LCS – longest common subsequence
  - WLCS – weighted LCS

MATERIALS – synthetic data

- Synthetic problem
  - Artificial web site map
  - Artificially generated clickstream data
- Sessions, similar to real ones
  - Average session length
  - Lower number of total site pages (30 pages)
- Protocol
  - User sessions generated according to site map
  - Generation of clickstream data
  - Separation process
- Separation: about 90%, perfect match
- Heuristic function for session length 10: on average 712 of 140.000 states



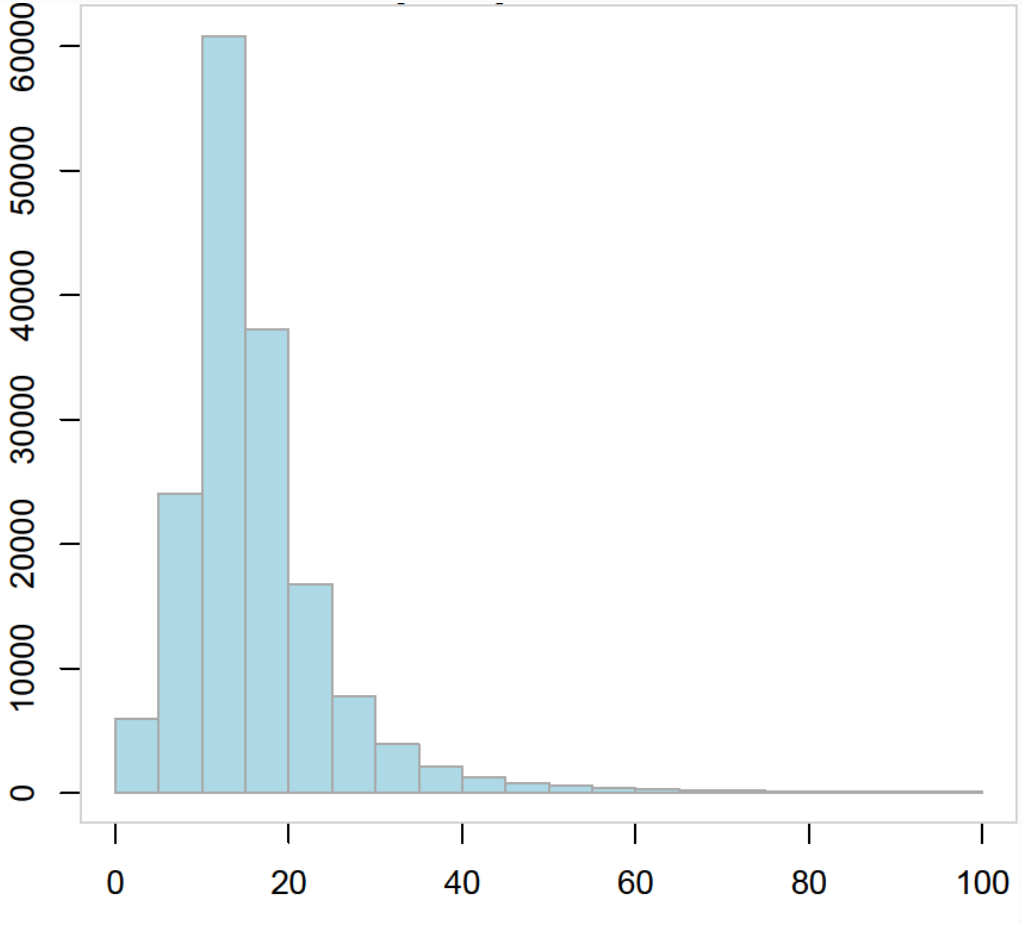
Real-world data

- Two real clickstream data sources
  - Student records information system
  - Web shop
- Considerably different types of clickstream data

Student records IS

- Approx. 300 different web pages, 160.000 validated user sessions
- Each state in MM corresponds to one web page
- Typical user paths well defined
- User has to log on (user identity is known)
- Easily identified entry point
- Server log files use basic CLF format
- Interleaved sessions: user with different concurrent user roles

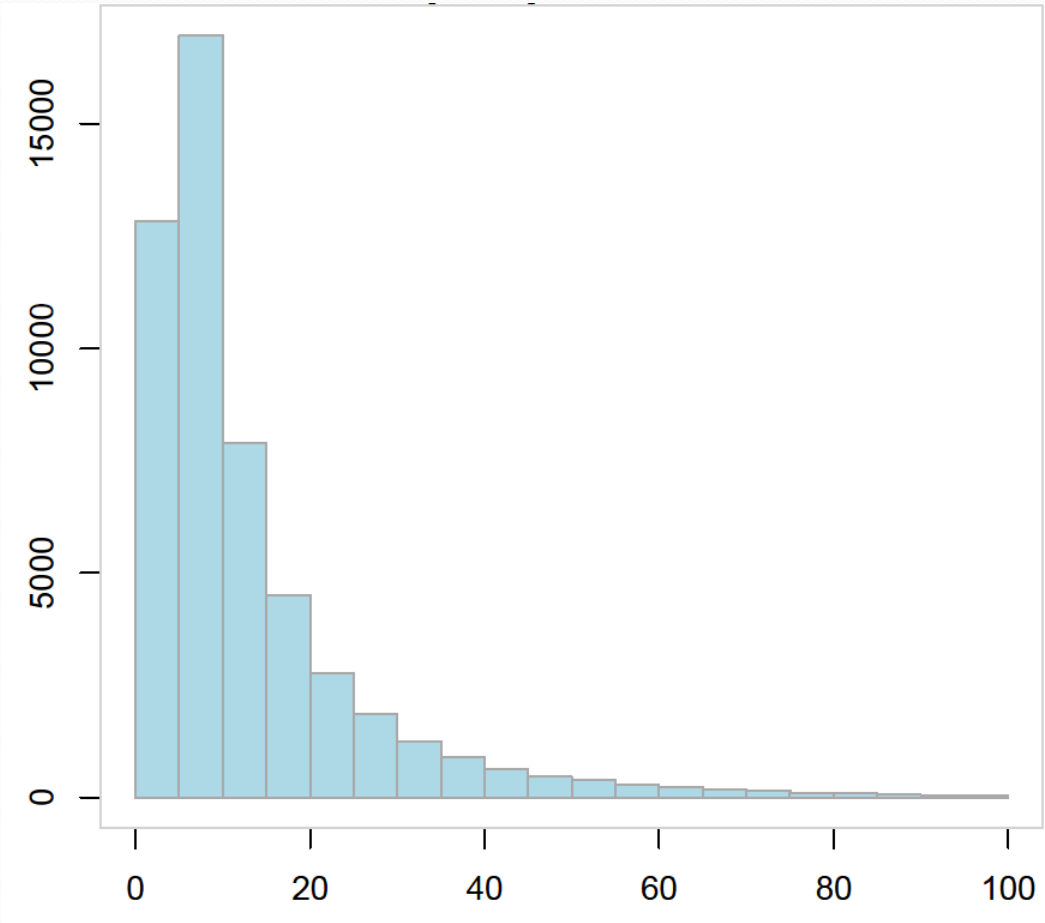
Session length for the student IS



Web shop

- Lots of application pages (tens of thousands) and users sessions (millions, 50.000 validated)
- Each state in MM corresponds to a group of pages
- Typical user path is not well defined
- Logon not required
  - User identity is not known
  - Logon only when purchase is made
- Entry point can be almost any page
- Pages are strongly linked (little use of site map)
- User login not required
- Hard to identify and eliminate Web robots

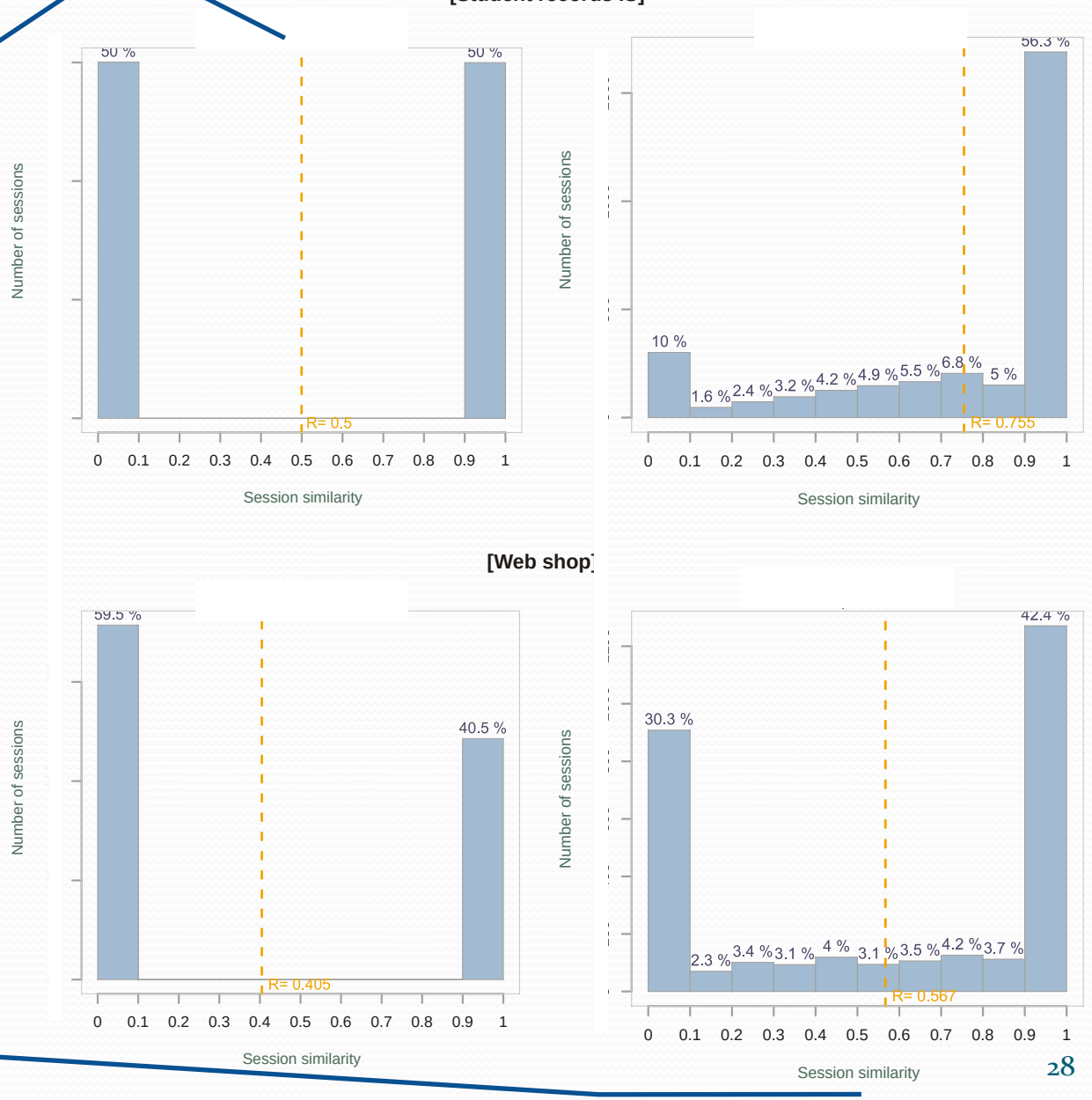
Session length for the web shop



Results

Perfect match

WCLS



Conclusion

- A new method for improved clickstream data pre-processing
- Data representation is based on first-order discrete MM
- Method
  - based on best-first heuristic search
  - tested on two real-world clickstream data sources
- Experiments on two datasets quite successful
  - promising results
  - can be used on any clickstream data source
  - independant of the number of consisting sessions

Lessons learned and further work

- More data for training (hundreds of thousands or millions of sessions)
- Better context help (semantic web?)
- Better utilization of available memory (SMA*)
- Estimate reliability of separation process
  - Probability
  - Number of searched states